４Ｕ－３ ゲノムデータベースにおける エントリの関連性検索

三村　徹　　諸岡　慎士　　山名　早人
早稲田大学理工学部情報学科

１．概要

近年、ヒトゲノム計画の進展に伴い生物情報が量産されるようになり、ゲノム情報を蓄積したデータベース（ゲノムデータベース）が作成された。

ゲノムデータベースは、エントリと呼ばれるテキスト形式を主とする要素の集合で構成されている**[図１]**。１つのエントリに記述される情報の範囲は、１つの遺伝子がエントリに相当するものもあれば、遺伝子の中のエキソンや断片であることもある。また、各エントリにはエントリＩＤ（またはアクセッション番号）と呼ばれる識別子が与えられている。

エントリは、フィールドと呼ばれる領域に分かれている。さらにフィールドは、タイトル、採取生物、塩基配列など、各情報に合わせて作られており、何の情報が記載されているか判るようにフィールド名が付けられている。

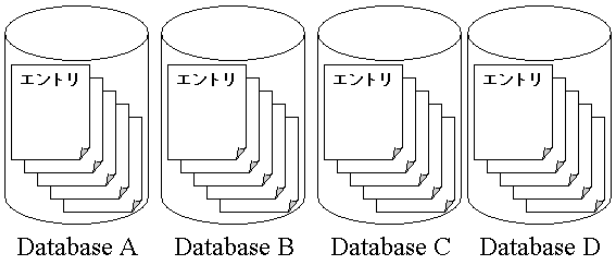

図1:ゲノムデータベースの構造

ゲノムデータベースの検索において、検索結果として得られる大量のエントリのリストから、検索要求（クエリ）と関連の強いエントリ群を絞り込むことが必要とされている**[1]**。関連があると言っても、利用者が重要視するフィールドにおいて、関連の強さを具体的な尺度を通じて知ることができなければ、関連性の意味や重要性は測れない。

本研究の目標は、利用者が重要視したフィールドにおける、関連の強いエントリ群のリストと、関連の強さを表すスコアを取得することである。検索対象のフィールドを複数選択し、各フィールドごとに関連を調べ、関連の強さに応じたスコアを与えるという手順で実現する。

そこで本システムでは、比較元のエントリ（クエリエントリ）に対し、そのエントリに関連する様々なエントリを、関連の強い順に取得する。その際に、利用者が重要視するフィールドに、重要性に応じた重みを与えて検索することで、エントリ全体としての関連性に方向性を持たせることができる。

２．手順

手順は、①準備として、ゲノムデータベースのエントリ構成を、検索に適したインデックスファイルに再構成する。②フィールドごとに関連性をスコアリングする。③エントリ全体の関連性を示すスコアに帰着させる。以上の３段階を経る。

２．１インデクシング

準備として、エントリをフィールドごとに分割し、単語のみを抽出する。エントリは、関連の対象をカスタマイズするのに適したデータ構造ではない。なぜならエントリには、箇条書きのフィールド、配列情報のフィールド、そして自然言語で記述されたフィールドのように、様々な記述形式が存在するからである**[２]**。

そこで、ゲノムデータベースのエントリ単位の構造を、関連性のスコアリングに適した構造に再構成する。手順としては、(1)フィールドの分割、(2)インデクシング、の２つの段階を経る **[図２]**。

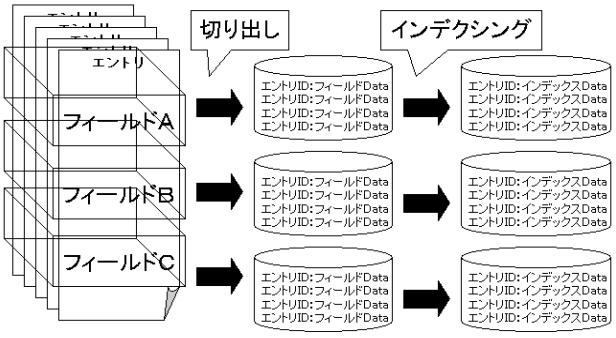

図2:インデクシング

(1)フィールドの分割では、エントリをフィールドごとに、１フィールドに対して１ファイルとなるように切り出し、フィールド名に従ってディレクトリに分ける。

(２)インデクシングでは、フィールドから単語のみを抽出する。

自然言語で書かれたフィールドに対しては、ApplePieParser などの形態素解析ツールを用いて、接続詞、前置詞などの関連性を示す単語になり得ないものを除外して、適切にインデクシングする必要がある。

Entry-Relevance Searches of Genome Database
Toru Mimura, Shinji Morooka, Hayato Yamana
Department of Information and Computer Science, School of Science and Engineering, Waseda Univ.

２．２フィールドレベルのスコアリング

次のステップとして、フィールドレベルの関連性を求め、関連性の強さに応じたスコアを与える。配列情報に関するフィールドには、既に BLAST[3] や FASTA[4] といった相同性検索ツールが存在する。しかし、配列の相同性を検索するだけであり、機能などにおける関連の強いエントリを検索するには至らない。

ここで用いるスコアリング手法は、比較元のフィールド（クエリ）と、全てのフィールドを比較して、両方のフィールドに出現している単語を、動的計画法によりカウントする方法である[図３]。関連性の強さを測るスコアリング手法は、何の関係も無いフィールドに比べ、同じ単語が出現しているフィールドの方が関連性は強いものとなっているはずである、という推測に基づいている。

	A	B	D	F	G	H	I	A	B
A	**1**	1	1	1	1	1	1	**2**	2
B	2	**3**	3	3	3	3	3	3	**4**
C	4	4	4	4	4	4	4	4	4
D	4	4	**5**	5	5	5	5	5	5
E	5	5	5	5	5	5	5	5	5
F	5	5	5	**6**	6	6	6	6	6
G	6	6	6	6	**7**	7	7	7	7

図3:フィールドレベルのスコアリング

図３のマトリクスにおいて、①縦の項目にクエリに出現する単語を並べ、横の項目に比較するフィールドに出現する単語を並べる。②単語の比較は、左上のマスから右に、１行ずつ進む。③単語が一致した時にカウントを１増やす。④右下の最後のマスに単語一致の回数が表れる。⑤単語一致の回数を、単語比較の回数で割ることで、両比較者全体の単語数を考慮したスコアが与えられ（図３の場合は **7/63**）、スコアは０～１で与えられる。このスコアをフィールドレベルのスコアと呼ぶことにする。

利用者がクエリエントリを決定したら、指定したフィールドに関してフィールドレベルのスコアリングを全エントリに渡って行う。

２．３スコアの統合

最後のステップとして、先に求めたフィールドレベルのスコアに重み付けし、統合することで、エントリ全体での関連性の高いエントリから順に求める手法に関して述べる。

ここで、エントリ全体でのスコア（フィールドレベルのスコアを統合したもの）を S、フィールド n に対するフィールドレベルのスコアを Fn、フィールド n に対する重みを Wn とすると、$0 \leqq Fn \leqq 1$, $0 \leqq Wn \leqq 1$, $\Sigma Wn = 1$ であり、エントリ全体でのスコアは $S = \Sigma Fn \times Wn$ で与えられる。フィールドが完全に一致した時のフィールドレベルのスコアは１となるので、同一のエントリに対してのエントリ全体のスコアは１となり、これがスコアの最高値である。逆に、同じ単語が１つも出現しないフィールドにおけるフィールドレベルのスコアは０となる。

このようにして全てのエントリについてエントリ全体でのスコアを求め、昇順に並び換えることで、検索は完了する。

３．考察

本研究において、ゲノムデータベースにおける新しい検索方法を提案した[図４]。この検索方法では、フィールドにおけるスコアリング手法をカスタマイズすることで、利用者の意図に合わせた検索を可能にする。

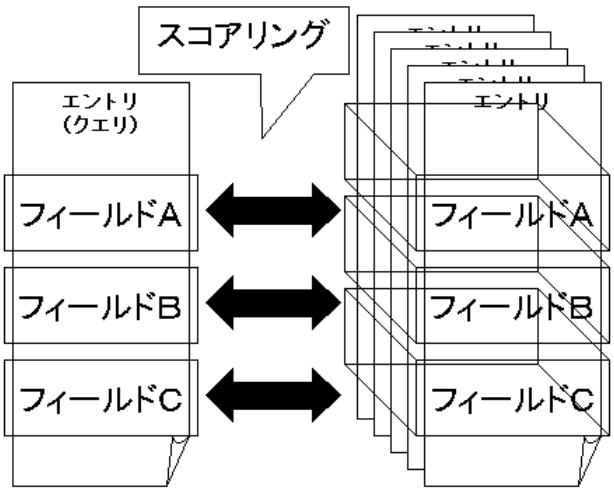

図4:検索コンセプト

例えば、塩基配列のフィールドに対して配列のサイズや塩基の割合をスコアリングの対象としたり、発表された年の近いものに高いスコアを与えることも可能である。

参考文献

[1]M.Kanehisa,:”A database for post-genome analysis” *Trends Genet*, 13, 375-376 (1997)
[2]M.Kanehisa, S.Goto, S.Kawashima, A.Nakaya:”The KEGG databases at GenomeNet” *Nucleic Acids Res*, 30, 42-46 (2002)
[3]S.F.Altschul, T.L.Madden, A.A.Achaeffer, J.Zhang, Z.Zhang, W.Miller, D.J.Lipman: ”Gapped BLAST and PSI-BLAST:a new generation of protain database search programs” *Nucleic Acid Res*, 25, 3389-3402 (1997)
[4]W.R.Pearson, D.J.Lipman:”Improved tools for biological sequence comparison” *Proc. Natl. Acad. Sci. USA*, 85, 2444-2448 (1988)